www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Identification of Jak-STAT signaling involvement in sarcoidosis severity via a novel microRNA-regulated peripheral blood mononuclear cell gene signature

Received: 14 September 2016
Accepted: 10 May 2017
Published online: 26 June 2017

Tong Zhou[1], Nancy Casanova[2], Nima Pouladi[3], Ting Wang[2], Yves Lussier[3], Kenneth S. Knox[2] & Joe G. N. Garcia[2]

Sarcoidosis is a granulomatous lung disorder of unknown cause. The majority of individuals with sarcoidosis spontaneously achieve full remission (uncomplicated sarcoidosis), however, ~20% of sarcoidosis-affected individuals experience progressive lung disease or cardiac and nervous system involvement (complicated sarcoidosis). We investigated peripheral blood mononuclear cell (PBMC) microRNA and protein-coding gene expression data from healthy controls and patients with uncomplicated or complicated sarcoidosis. We identified 46 microRNAs and 1,559 genes that were differentially expressed across a continuum of sarcoidosis severity (healthy control → uncomplicated sarcoidosis → complicated sarcoidosis). A total of 19 microRNA-mRNA regulatory pairs were identified within these deregulated microRNAs and mRNAs, which consisted of 17 unique protein-coding genes yielding a 17-gene signature. Pathway analysis of the 17-gene signature revealed Jak-STAT signaling pathway as the most significantly represented pathway. A severity score was assigned to each patient based on the expression of the 17-gene signature and a significant increasing trend in the severity score was observed from healthy control, to uncomplicated sarcoidosis, and finally to complicated sarcoidosis. In addition, this microRNA-regulated gene signature differentiates sarcoidosis patients from healthy controls in independent validation cohorts. Our study suggests that PBMC gene expression is useful in diagnosis of sarcoidosis.

Sarcoidosis remains a systemic inflammatory disease of unknown etiology with an unpredictable course that affects all races and ethnicities that is characterized by the presence of non-caseating epithelioid granulomas in one or multiple organs. The disorder is extremely heterogeneous with more than 50% of sarcoidosis patients experience remission within 3 years after diagnosis, and over 66% of patients experiencing remission within 10 years[1–3] (uncomplicated sarcoidosis), whereas a significant percentage of patients with sarcoidosis develop granulomatous involvement of other vital organs with progressive disease (complicated sarcoidosis). Lung involvement is commonly manifested as bilateral hilar lymphadenopathy and pulmonary infiltration with more severe cases developing pulmonary fibrosis. Cardiac and neurologic involvement are also associated with significant morbidity and death[3, 4]. This broad spectrum of clinical manifestations makes diagnosis challenging and prolonged.

Clearly validated biomarkers that can accurately sub-phenotype patients with sarcoidosis are of enormous utility as they may identify subjects at increased risk of complicated sarcoidosis and allow for the delivery of targeted therapies to prevent progressive multi-organ deterioration. Unfortunately, blood, bronchoalveolar lavage fluid, exhaled breath concentrate, and cerebrospinal fluid have all been assessed as sarcoidosis biomarkers by multiple methodologies, including enzyme-linked immunosorbent assay and proteomic analysis, but with

[1]Department of Physiology and Cell Biology, University of Nevada, Reno School of Medicine, Reno, NV, 89557, USA. [2]Division of Pulmonary, Allergy, Critical Care, and Sleep Medicine, Department of Medicine, University of Arizona Health Sciences, Tucson, AZ, 78721, USA. [3]Center for Bioinformatics and Biostatistics, University of Arizona Health Sciences, Tucson, AZ, 78721, USA. Tong Zhou and Nancy Casanova contributed equally to this work. Correspondence and requests for materials should be addressed to J.G.N.G. (email: skipgarcia@email.arizona.edu)

limited success. For example, angiotensin converting enzyme (ACE) is the most commonly studied biomarker but is non-specific and does not accurately correlate with severity[5, 6]. Acute phase reactants like serum amyloid (SAA) levels are significantly higher in active disease similarly to ACE[7], but likewise suffers from low specificity and sensitivity. HLA-DR allele and cytokines such as tumor necrosis factors (TNF-α, TNF-β) are overexpressed in sarcoidosis as well and correlated with fibrosis development[8]. No single biomarker has proven to exhibit significant sensitivity and specificity to be recommended as a monitoring and prognostic tool for standard clinical use.

As a consequence, we and others have proposed the integration of blood gene expression profiling as an opportunity to explore potential useful molecular signatures in sarcoidosis[9–11], particularly relevant in assessment of personal risk for developing complicated sarcoidosis[10]. For example, Koth *et al.* analyzed the transcriptomic gene expression data from whole blood of sarcoidosis patients and accordingly built a classifier using the gene expression data, which distinguished sarcoidosis patients from healthy controls in an external validation cohort with a fairly good accuracy[9]. Recently, we investigated genome-wide gene expression in peripheral blood mononuclear cell (PBMC) in sarcoidosis patients and identified a 20-gene signature, which distinguishes uncomplicated sarcoidosis from complicated sarcoidosis[10]. This gene signature also exhibited a substantial predictive power when classifying sarcoidosis patients from healthy controls in two external cohorts.

In this study, we proposed to incorporate microRNA (miRNA) expression information with protein-coding gene expression data to identify sarcoidosis biomarkers. Firstly, we identified a list of differentially expressed miRNAs in PBMCs from patients with sarcoidosis. Secondly, we searched for a list of target protein-coding genes for the deregulated miRNAs and based on these target genes, developed a novel miRNA-regulated peripheral blood gene signature that accurately differentiates sarcoidosis patients from healthy controls and complicated sarcoidosis patients from uncomplicated ones. Pathway analysis of the gene signature revealed Jak-STAT signaling pathway as the most significantly represented pathway. A severity score, based on the expression of the gene signature, exhibited a significantly increasing trend from healthy control, to uncomplicated sarcoidosis, and finally to complicated sarcoidosis. These studies suggest that PBMC gene expression is useful in diagnosis of sarcoidosis, and more importantly, in the identification of patients with complicated sarcoidosis.

Results

Identifying miRNAs differentially expressed in sarcoidosis PBMCs.

To determine the differentially expressed miRNAs in sarcoidosis PBMCs, we analyzed the PBMC miRNA expression data from 35 healthy controls, 17 patients with uncomplicated sarcoidosis, and 13 patients with complicated sarcoidosis. Spearman's rank correlation test was used to identify the miRNAs that were differentially expressed with sarcoidosis severity (healthy control → uncomplicated sarcoidosis → complicated sarcoidosis). In total, we identified 46 miRNAs (adjusted $P < 0.05$) that were differentially expressed with severity (Supplementary Table S1). Nineteen out of the 46 miRNAs showed positive correlation between expression and severity while 27 miRNAs showed negative correlation (Supplementary Table S1).

Identifying protein-coding genes differentially expressed in sarcoidosis PBMCs.

To determine the differentially expressed protein-coding genes in sarcoidosis PBMCs, we analyzed the gene expression pattern from 35 healthy controls, 17 patients with uncomplicated sarcoidosis, and 22 patients with complicated sarcoidosis (Gene Expression Omnibus [GEO][12] accession: GSE37912). Spearman's rank correlation test was conducted between gene expression and sarcoidosis severity. In total, 1,559 genes showed significant correlation (adjusted $P < 0.0005$), among which 340 genes showed positive correlation between gene expression and severity while the expression of 1,219 genes was negatively correlated with severity (Supplementary Table S2).

In silico prediction of miRNA-mRNA binding pairs.

We searched for the target genes of the differentially expressed miRNAs, using the *in silico* prediction provided by microrna.org[13], with filtering based on a stringent mirSVR score cutoff of -1.2 (this value represents the top 5% of mirSVR scores) to select the top miRNA-mRNA binding pairs[14]. We intersected these predicted binding targets against the differentially expressed protein-coding genes revealing a total of 425 miRNA-mRNA pairs. Among these miRNA-mRNA pairs, we only retained the pairs that demonstrated a significant negative correlation (adjusted $P < 0.005$) between miRNA and mRNA expression. This step yielded 19 miRNA-mRNA pairs consisting of eight unique miRNAs and 17 unique protein-coding genes (Table 1 and Supplementary Fig. S1). We designated the eight miRNAs as an 8-miRNA signature (Table 2), among which seven miRNAs were upregulated in complicated sarcoidosis with only one downregulated miRNA (Fig. 1). The list of 17 protein-coding genes was termed a 17-gene signature (Table 3) and contained only one gene upregulated in complicated sarcoidosis whereas the other 16 genes were downregulated (Fig. 2A). We next searched the enriched Gene Ontology (GO) biological process terms[15] for the 17-gene signature and found that the 17-gene signature is significantly associated with "JAK-STAT cascade" (Fig. 2B). Pathway analysis based on Kyoto Encyclopedia of Genes and Genomes (KEGG)[16] database also confirmed that the top KEGG pathway associated with the 17-gene signature is "Jak-STAT signaling pathway" (Fig. 2C).

Classification power of the 8-miRNA signature.

We first tested the classification power of the 8-miRNA signature in our discovery cohort. A severity score (S_{miRNA}) was assigned to each patient based on the expression of the 8-miRNA signature, which is a linear combination of the miRNA expression values weighted by direction of differential expression (See Methods for details). Higher S_{miRNA} implies more severe sarcoidosis. As expected, a significant positive correlation (Spearman's rank correlation test: $\rho = 0.651$ and $P = 4.4 \times 10^{-9}$) was identified between S_{miRNA} and sarcoidosis severity in our discovery cohort (Supplementary Fig. S2).

We next investigated the classification power of the 8-miRNA signature in a validation dataset from Germany (GEO accession: GSE34608)[17], which contains the whole blood gene expression data for eight healthy controls and eight patients with sarcoidosis (Germany cohort). The 8-miRNA signature based severity score of sarcoidosis

miRNA	Target gene	mirSVR score	ρ	Adjusted *P*
miR-23a	*EFHA2*	−1.260	−0.589	3.42×10^{-5}
miR-23a	*GALNT12*	−1.264	−0.480	1.78×10^{-3}
miR-23a	*SATB1*	−1.342	−0.454	3.75×10^{-3}
miR-23a	*STAT4*	−1.276	−0.475	2.12×10^{-3}
miR-23a	*TMEM263*	−1.201	−0.459	3.26×10^{-3}
miR-23b	*EFHA2*	−1.260	−0.531	3.49×10^{-4}
miR-23b	*GALNT12*	−1.264	−0.438	5.75×10^{-3}
miR-30c	*ITGA6*	−1.203	−0.433	6.46×10^{-3}
miR-93	*FIGNL1*	−1.231	−0.463	2.91×10^{-3}
miR-93	*MBTPS1*	−1.228	−0.533	3.23×10^{-4}
miR-93	*MTERFD2*	−1.231	−0.540	2.51×10^{-4}
miR-93	*URI1*	−1.273	−0.468	2.57×10^{-3}
miR-93	*ZFYVE9*	−1.332	−0.511	7.03×10^{-4}
miR-143	*ATP10A*	−1.333	−0.443	5.05×10^{-3}
miR-185	*SORCS3*	−1.215	−0.437	5.79×10^{-3}
*miR-196a**	*ADORA3*	−1.270	−0.472	2.30×10^{-3}
miR-223	*CBLB*	−1.252	−0.447	4.50×10^{-3}
miR-223	*ERCC6L2*	−1.296	−0.446	4.63×10^{-3}
miR-223	*IL6ST*	−1.340	−0.432	6.53×10^{-3}

Table 1. Sarcoidosis-related miRNA and target gene pairs. Note – ρ is the Spearman's rank correlation coefficient. *P*-values were calculated by Spearman's rank correlation test between miRNA and target gene expression levels and adjusted by Benjamini & Hochberg procedure.

miRNA	ρ	Adjusted *P*
miR-23a	0.377	1.91×10^{-2}
miR-23b	0.405	1.33×10^{-2}
miR-30c	0.344	3.44×10^{-2}
miR-93	0.386	1.78×10^{-2}
miR-143	0.378	1.91×10^{-2}
miR-185	0.363	2.43×10^{-2}
*miR-196a**	−0.393	1.60×10^{-2}
miR-223	0.346	3.44×10^{-2}

Table 2. 8-miRNA signature in Sarcoidosis. Note – ρ is the Spearman's rank correlation coefficient. *P*-values were calculated by Spearman's rank correlation test between miRNA expression level and sarcoidosis severity and adjusted by Benjamini & Hochberg procedure.

patients is significantly higher than that of controls in the Germany cohort (t-test: $P = 3.4 \times 10^{-6}$) (Fig. 3A). The areas under the receiver operating characteristic (ROC) curve (*AUC*) is 1.000 in this cohort (Fig. 3B), which suggests a fairly good classification accuracy of the 8-miRNA signature. Principal component analysis on the 8-miRNA expression also indicates that the sarcoidosis patients can be clearly distinguished from the health controls (Fig. 3C).

Classification power of the 17-gene signature. Similar to the strategy employed for the 8-miRNA signature, we assigned a severity score (S_{gene}) to each patient based on the expression of the 17-gene signature, which is a linear combination of the protein-coding gene expression values weighted by direction of differential expression (see Methods for details). Patients with more severe sarcoidosis are expected to have a higher S_{gene}. Not surprisingly, we found a significant positive correlation (Spearman's rank correlation test: $\rho = 0.615$ and $P = 5.6 \times 10^{-9}$) between S_{gene} and sarcoidosis severity in our discovery cohort (Supplementary Fig. S3).

We next validated the predictive power of the 17-gene signature in two independent blood gene expression datasets. One dataset (GEO accession: GSE19314)[9] is from University of California, San Francisco that contains 20 healthy controls and 40 sarcoidosis patients (UCSF cohort). The other dataset (GEO accession: GSE18781)[18] is from Oregon Health Sciences University and includes 25 healthy controls and 12 sarcoidosis patients (Oregon cohort). The 17-gene signature-based severity score of sarcoidosis patients is significantly higher than that of controls in both the validation cohorts (t-test: $P = 1.5 \times 10^{-8}$ for the UCSF cohort and $P = 3.8 \times 10^{-5}$ for the Oregon cohort) (Fig. 4A). The *AUC* values of classification are 0.876 and 0.913 in the UCSF and Oregon cohorts, respectively (Fig. 4B). Principal component analysis on the expression of the 17-gene signature indicates that the sarcoidosis patients can be well distinguished from the health controls (Supplementary Fig. S4).

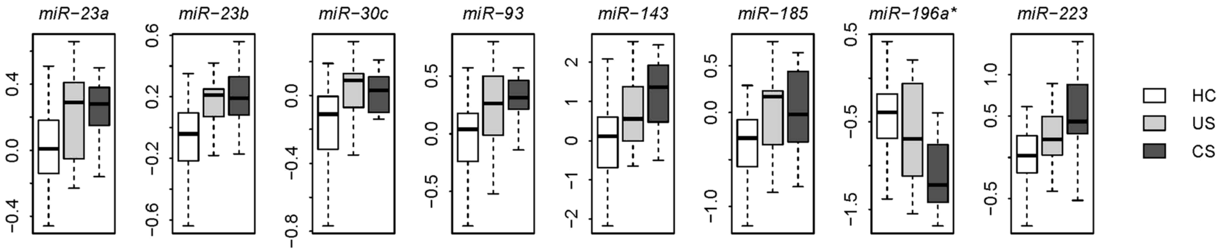

Figure 1. 8-miRNA signature in sarcoidosis. The eight miRNAs were differentially expressed with the severity of sarcoidosis. Y-axis indicates the miRNA expression level. HC: healthy controls; US: uncomplicated sarcoidosis; CS: complicated sarcoidosis.

Gene symbol	Gene title	ρ	Adjusted *P*
ADORA3	adenosine A3 receptor	0.463	3.04×10^{-4}
ATP10A	ATPase, class V, type 10A	−0.541	3.58×10^{-5}
CBLB	Cas-Br-M (murine) ecotropic retroviral transforming sequence b	−0.483	1.75×10^{-4}
EFHA2	EF-hand domain family, member A2	−0.505	9.48×10^{-5}
ERCC6L2	excision repair cross-complementation group 6-like 2	−0.508	8.92×10^{-5}
FIGNL1	fidgetin-like 1	−0.565	1.77×10^{-5}
GALNT12	polypeptide N-acetylgalactosaminyltransferase 12	−0.688	7.40×10^{-8}
IL6ST	interleukin 6 signal transducer (gp130, oncostatin M receptor)	−0.470	2.52×10^{-4}
ITGA6	integrin, alpha 6	−0.510	8.18×10^{-5}
MBTPS1	membrane-bound transcription factor peptidase, site 1	−0.512	7.89×10^{-5}
MTERFD2	MTERF domain containing 2	−0.533	4.31×10^{-5}
SATB1	SATB homeobox 1	−0.457	3.65×10^{-4}
SORCS3	sortilin-related VPS10 domain containing receptor 3	−0.457	3.62×10^{-4}
STAT4	signal transducer and activator of transcription 4	−0.500	1.11×10^{-4}
TMEM263	transmembrane protein 263	−0.463	3.06×10^{-4}
URI1	URI1, prefoldin-like chaperone	−0.503	1.01×10^{-4}
ZFYVE9	zinc finger, FYVE domain containing 9	−0.478	2.01×10^{-4}

Table 3. 17-Gene Signature in Sarcoidosis. Note – ρ is the Spearman's rank correlation coefficient. *P*-values were calculated by Spearman's rank correlation test between protein-coding gene expression level and sarcoidosis severity and adjusted by Benjamini & Hochberg procedure.

To understand the discriminative power of the 17-gene signature in distinguishing sarcoidosis from other lung diseases, we collected two more blood transcriptomic datasets (GEO accession: GSE42826 and GSE4257), the first consisted of 52 healthy controls, 25 sarcoidosis patients, 11 tuberculosis (TB) patients, six pneumonia patients, and eight lung cancer patients. TB, pneumonia patients, lung cancer were considered as non-sarcoidosis lung disorders. The second dataset comprised of 123 healthy controls, 87 and 47 cases with COPD and sarcoidosis respectively. Principal component analysis on the expression of the 17-gene signature in this dataset indicates that the first principal component significantly differentiates the sarcoidosis patients from the non-sarcoidosis lung disorders in both datasets (t-test: $P = 1.9 \times 10^{-2}$ and 2.09×10^{-5}) (Supplementary Figs S5 and S6), which suggests that the 17-gene signature is able to distinguish between sarcoidosis and non-sarcoidosis lung disorders. To compare the effect of systemic corticosteroids on our 17-gene signature, we also surveyed a gene expression data set (GSE67255) on 18 healthy subjects before and 24 hours after the administration of glucocorticoids[19]. The differential expression was performed using one-way repeated measure ANOVA. None of the 17 genes in our signature were significantly affected ($P < 0.5$).

A computational study by Venet *et al.* suggests that most published gene signatures are not significantly better than random gene sets of identical size that are randomly picked up from human genome[20]. To address this issue, resampling test was applied. We obtained 1,000 random gene signatures by randomly selecting 17 genes from human genome. For each random gene set, we calculated the *AUC* values for both the UCSF and Oregon cohorts. The sum of *AUC* in both cohorts was recorded for each random gene signature, which measured the classification power of the random gene set. Our alternative hypothesis is that the sum of *AUC* of our 17-gene signature should be more positive than expected by chance if the classification power of the 17-gene signature is significantly better than the random gene signatures. By resampling test, we found that we could reject the null hypothesis that the classification power of the 17-gene signature is by chance. The sum of *AUC* of our 17-gene signature is significantly larger than that of random gene signatures (Right-tailed: $P = 0.001$) (Fig. 4C).

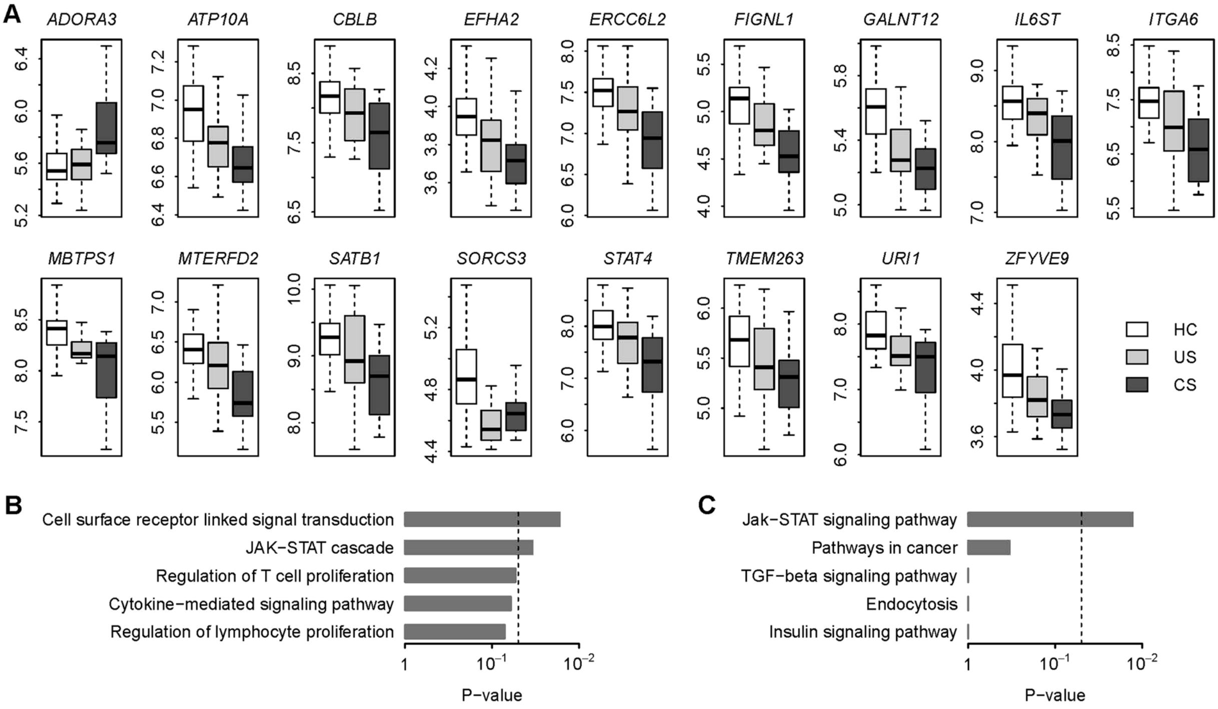

Figure 2. 17-gene signature in sarcoidosis. (**A**) The 17 protein-coding genes were differentially expressed with the severity of sarcoidosis. Y-axis indicates the gene expression level. HC: healthy controls; US: uncomplicated sarcoidosis; CS: complicated sarcoidosis. (**B**) The top five GO biological process terms associated with the 17-gene signature. The *P*-values were calculated by Fisher's exact test. The dash line denotes the significance level of 0.05. (**C**) The top five KEGG pathway terms associated with the 17-gene signature. The *P*-values were calculated by Fisher's exact test. The dash line denotes the significance level of 0.05.

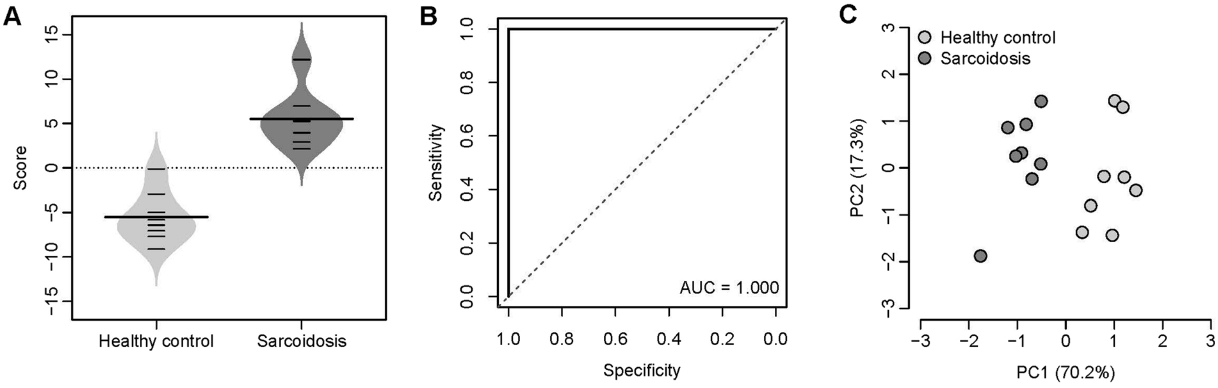

Figure 3. The performance of the 8-miRNA signature in the sarcoidosis validation cohort. (**A**) The 8-miRNA signature based severity score differentiates the sarcoidosis patients from the healthy controls in the Berlin cohort. The violin plot indicates the distribution of the severity score in each category. (**B**) The ROC curve of the 8-miRNA signature in classifying the subjects in the Berlin cohort. The AUC is equal to one. (**C**) Principal component analysis on the expression of the 8-miRNA signature. X-axis: the first principal component with eigenvalue; Y-axis: the second principal component with eigenvalue.

We next addressed whether the classification power of the 17-gene signature is superior to the other genes that are related to sarcoidosis. To answer this question, we conducted a second resampling test. We limited the resampling pool to the genes that were differentially expressed with sarcoidosis severity (1,559 genes listed in Supplementary Table S2) and defined these genes as sarcoidosis related. We then randomly selected 17 genes from the pool of sarcoidosis related genes and tested the predictive power of the random gene signature. We repeated this randomization procedure for 1,000 times. The performance of the random gene signature was quantified by the sum of *AUC* in both validation cohorts. Figure 4C demonstrates that the classification power of the 17-gene signature is significantly better than that of 1,000 resampled sarcoidosis related gene signatures (Right-tailed: $P = 0.019$).

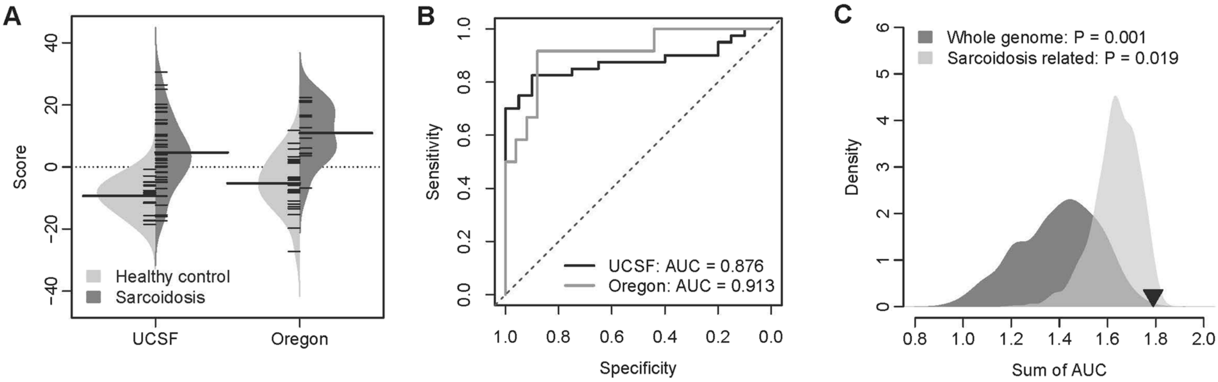

Figure 4. The performance of the 17-gene signature in the validation cohorts. (**A**) The 17-gene signature based severity score differentiates the sarcoidosis patients from the healthy controls in the UCSF and Oregon cohorts. The violin plot indicates the distribution of the severity score in each category. (**B**) The ROC curves of the 17-gene signature in classifying the subjects in the UCSF and Oregon cohorts. (**C**) Superior predictive power of the 17-gene signature compared with random gene set. The dark grey area shows the distribution of the sum of AUC (both the validation cohorts) for the 1,000 resampled gene signatures (with the identical size as the 17-gene signature) randomly picked up from human genome. The light grey area shows the distribution of the sum of AUC for the 1,000 resampled gene signatures randomly selected from the pool of the sarcoidosis related genes. The black triangle stands for the sum of AUC of the 17-gene signature. Right-tailed *P*-values of the sampling distribution were calculated.

Discussion

Sarcoidosis remains a multifactorial, complex, and challenging systemic disease, whose unpredictable course mandates the urgent application of novel approaches, such as genome-based expression profiling, to generate personalized risk assessment tools to diagnose, monitor disease progression, and guide therapeutic management of those with this affliction. Here, we stratified sarcoidosis cases into two different categories: uncomplicated and complicated sarcoidosis. Uncomplicated cases presented remission, while complicated cases exhibited significant organ impairment. The accurate identification of patients at risk for complicated sarcoidosis is a clinical challenge. Biomarkers or molecular signatures emerging from new technologies such as PBMC miRNA/gene expression analysis represent an opportunity to change the routine clinical care of sarcoidosis. In this study, we investigated the PBMC miRNA and protein-coding gene expression data from both healthy controls and patients with sarcoidosis. Using these data, we identified a molecular signature consisting of 17 protein-coding genes for diagnostic purpose in sarcoidosis. A severity score was assigned to each human subject based on the expression of the 17-gene signature. A higher score suggested a higher likelihood of complicated sarcoidosis. In our discovery cohort, we observed a significant increasing trend in the severity score from healthy control, uncomplicated sarcoidosis, to complicated sarcoidosis. We also demonstrated that this miRNA-regulated gene signature can differentiate sarcoidosis patients from healthy controls in independent validation cohorts.

We initially identified a 8-miRNA signature (7 upregulated and 1 downregulated miRNAs in complicated sarcoidosis) that potentially relate to sarcoidosis severity, including miR-23a, miR-23b, miR-30c, miR-185, and miR-223, now recognized to be related to pulmonary hypertension[21–25] and lung cancer[26–29]. In a previous study, PBMC miRNA expression profiles were compared between sarcoidosis patients and healthy controls[30]. However, we failed to identify any single miRNA within our 8-miRNA signature that overlapped with the reported list of differentially expressed miRNAs[30]. This lack of overlap may reflect the difference in methodology and objective between the two studies: we mainly focused on complicated sarcoidosis and correlation test was applied to find the miRNAs associated with sarcoidosis severity. Based on the expression of the 8-miRNA signature, we developed a scoring system. This severity score reflected the extent and severity of sarcoidosis. We demonstrated that this miRNA-based score differentiates sarcoidosis patients in both discovery and validation cohorts.

We also explored the utility of a protein-coding gene expression signature in sarcoidosis. We identified the 17-gene signature, in which the genes were potentially regulated by the miRNAs within the 8-miRNA signature. In agreement with prior reported studies, the 17-gene signature was enriched in Jak-STAT signaling pathway[9, 10, 18]. Jak-STAT signaling pathway is an intracellular cascade initiated in response to cytokine signaling. Several genes in Jak-STAT pathways have already been associated with sarcoidosis, such as *IL15*[31], *IL23R*[32], and *STAT1*[18]. According to the current knowledge of the molecular mechanism leading to sarcoidosis, genes involved in T-cell receptor selection, T-cell activation and apoptosis, and cytokine regulation are also highly associated with the pathogenesis of sarcoidosis[9, 10, 33]. In our 17-gene signature, *CBLB* and *IL6ST* are known to be involved in T-cell receptor signaling and cytokine-cytokine receptor interaction, respectively, according to the definition in KEGG database. However, we failed to identify a single gene within our 17-gene signature that overlapped with the previously published sarcoidosis blood gene signatures by Koth *et al.*[9] and Zhou *et al.*[10], respectively.

A published bioinformatical study by Venet *et al.* suggests that resampling test should be used to evaluate the predictive power of a given gene signature instead of focusing on nominal *P*-value, because most the published signatures are not significantly better than the randomized gene signatures of identical size[20]. Using resampling tests, we found that the classification power of the 17-gene signature is significantly better than that of the random gene sets selected from human genome.

In contrast to the published sarcoidosis gene signatures derived from whole genome screening by Koth *et al.*[9] and Zhou *et al.*[10], the 17-gene signature was developed based on miRNA expression information and predefined miRNA-gene interactions. Statistically-derived gene signatures by whole genome screening are often highly accurate in the patient data sets from which they were identified, yet most of them have not been validated as useful clinical tools. Using resampling test, we demonstrate that the miRNA-regulated 17-gene signature performs even better than the random gene sets selected from the pool of sarcoidosis related genes, which suggests that incorporation miRNA expression information into gene signature identification may significantly decrease false positive rate when developing gene signatures for complex human diseases.

Despite multiple efforts to identify concordant DE-miRNA in sarcoidosis, it is recognized that patterns of miRNA expression differ between BAL cells, lung tissue, lymph nodes and PBMCs suggesting organ specific regulation. In our initial 8-miRNa signature we identified miR-23a, miR-23b, miR-30c, miR-185, and miR-223 related to sarcoidosis severity. Kiszalkiewicz *et al.* recently published overexpression of miRNas involved in angiogenesis, miRNA -27b, miR-192 and miR-221 in BALF cells from patients with acute sarcoidosis, none of these were significantly overexpressed in PBMCs[34]. Crouser *et al.* also profiled DE-miRNA in lung and lymph tissue and observed no overlap in miRNA expression in PBMC relative to diseased lung tissue; however TGF Fβ/WNT was a commonly regulated pathway[30]. Jazwa *et al.*, suggested miRNA-34 regulates *SIRT1* and the expression of IFN- γ in sarcoidosis[34, 35].

In summary, we derived a molecular signature consisting of 17 protein-coding genes, which are potentially regulated by deregulated miRNA in sarcoidosis. This signature can be independently used as potential novel molecular markers for differentiating patients with sarcoidosis, especially for distinguishing the patients with risk of complicated sarcoidosis. Although gene expression variation in PBMCs may not necessarily reflect the dynamics of the tissue microenvironment in sarcoidosis patients, PBMC gene expression information is useful in diagnosis of sarcoidosis, and more importantly, in the identification of patients with complicated sarcoidosis.

Methods

PBMC samples.

The study was approved by the University of Arizona Institutional Review Board with written informed consent obtained from all subjects, and was performed in accordance with the principles in the Declaration of Helsinki. We defined complicated sarcoidosis as cardiac sarcoidosis (e.g. ventricular arrhythmias)[36], neurologic sarcoidosis (e.g. evidence of hyperdense MRI lesions)[37], or severe pulmonary sarcoidosis (forced vital capacity < 50%) (Supplementary Table S3). The diagnosis of sarcoidosis was based on established joint international criteria[38]. For our discovery cohort, PBMC samples were collected from 35 healthy controls and 39 sarcoidosis patients (17 patients with uncomplicated sarcoidosis and 22 patients with complicated sarcoidosis). The healthy controls were collected in a matching procedure for both age and gender of the patients. As for the sarcoidosis patients, we didn't find significant difference in age (t-test: P > 0.05) and gender (χ2-test: P > 0.05) between uncomplicated and complicated cases. Patient characteristics and treatment information are summarized in Supplementary Table S4.

High-throughput miRNA expression data.

For the discovery cohort, we profiled the PBMC miRNA expression level for 35 healthy controls, 17 patients with uncomplicated sarcoidosis, and 13 patients with complicated sarcoidosis, using Exiqon miRCURY LNA Array v11.0 (Exiqon, Inc., Denmark). Total RNA was isolated from PBMCs according to manufacturer's protocol. Array hybridization was performed by Exiqon with the quantified signals background corrected using *normexp* with offset value 10 based on a convolution model[39] and normalized using the global Lowess regression algorithm. Only the miRNAs being present in at least two third of total samples were further analyzed. The miRNA expression data of the validation cohort, the Germany cohort, were obtained from the GEO database (GEO accession: GSE34608), which were based on Agilent-019118 Human miRNA Microarray 2.0 G4470B.

High-throughput protein-coding gene expression data.

The protein-coding gene expression data of the discovery cohort are available from the GEO database (GEO accession: GSE37912)[10], which were based on Affymetrix Human Exon 1.0ST Array. Briefly, gene expression data were summarized using robust multi-array average (RMA) algorithm[40] embedded in the Affymetrix Power Tools v.1.12.0. Adjustment for possible batch effect was conducted by COMBAT[41]. We consider a transcript cluster to be reliably expressed in these samples if the Affymetrix implemented DABG (detection above ground)[42] *P*-value was less than 0.01 in at least two third of total samples. The gene expression data of the UCSF (GEO accession: GSE19314)[9] and Oregon cohorts (GEO accession: GSE18781)[18] were downloaded from the GEO database, which were based on Affymetrix Human Genome U133 Plus 2.0 Array.

Severity score.

8-miRNA and 17-gene based severity scores were calculated for each human subject, respectively. 8-miRNA based severity score (S_{miRNA}) is a linear combination of expression values of the miRNAs within the 8-miRNA signature. The formula is shown below:

$$S_{miRNA} = \sum_{i=1}^{N_{miRNA}} \mathrm{sgn}(\rho_i^{miRNA})(e_i^{miRNA} - \mu_i^{miRNA})/\tau_i^{miRNA} \tag{1}$$

Here, N_{miRNA} is the number of miRNAs in the 8-miRNA signature; ρ_i^{miRNA} is the Spearman's rank correlation coefficient of miRNA *i* (as shown in Table 2); e_i^{miRNA} is the expression level of miRNA *i*; μ_i^{miRNA} and τ_i^{miRNA} are the mean and standard deviation of the expression values for m*i*RNA *i* across all samples, respectively; and "sgn" denotes the sign function. 17-gene based severity score (S_{gene}) is a linear combination of expression values of the genes within the 17-gene signature[43–45]. The formula is shown below:

$$S_{gene} = \sum_{i=1}^{N_{gene}} \text{sgn}(\rho_i^{gene})(e_i^{gene} - \mu_i^{gene})/\tau_i^{gene} \tag{2}$$

Here, N_{gene} is the number of protein-coding genes in the 17-gene signature; ρ_i^{gene} is the Spearman's rank correlation coefficient of gene *i* (as shown in Table 3); e_i^{gene} is the expression level of gene *i*; μ_i^{gene} and τ_i^{gene} are the mean and standard deviation of the expression values for gene *i* across all samples, respectively; and "sgn" denotes the sign function.

References

1. Nunes, H., Bouvry, D., Soler, P. & Valeyre, D. Sarcoidosis. *Orphanet J Rare Dis* **2**, 46, doi:10.1186/1750-1172-2-46 (2007).
2. National Heart, L. A. B. I. *What Is Sarcoidosis?* https://www.nhlbi.nih.gov/health/health-topics/topics/sarc/.
3. Iannuzzi, M. C., Rybicki, B. A. & Teirstein, A. S. Sarcoidosis. *N Engl J Med* **357**, 2153-2165, doi:357/21/2153 10.1056/NEJMra071714 (2007).
4. Newman, L. S., Rose, C. S. & Maier, L. A. Sarcoidosis. *N Engl J Med* **336**, 1224–1234 (1997).
5. Baudin, B. [Angiotensin I-converting enzyme (ACE) for sarcoidosis diagnosis]. *Pathologie-biologie* **53**, 183–188, doi:10.1016/j.patbio.2004.09.003 (2005).
6. Ahmadzai, H. *et al.* Measurement of neopterin, TGF-beta(1) and ACE in the exhaled breath condensate of patients with sarcoidosis. *Journal of Breath Research* **7**, doi:10.1088/1752-7155/7/4/046003 (2013).
7. Gungor, S. *et al.* Conventional markers in determination of activity of sarcoidosis. *Int Immunopharmacol* **25**, 174–179, doi:10.1016/j.intimp.2015.01.015 (2015).
8. Mortaz, E. *et al.* Association of serum TNF-alpha, IL-8 and free light chain with HLA-DR B alleles expression in pulmonary and extra-pulmonary sarcoidosis. *Journal of inflammation (London, England)* **12**, 21, doi:10.1186/s12950-015-0066-3 (2015).
9. Koth, L. L. *et al.* Sarcoidosis blood transcriptome reflects lung inflammation and overlaps with tuberculosis. *Am J Respir Crit Care Med* **184**, 1153–1163, doi:10.1164/rccm.201106-1143OC201106-1143OC (2011).
10. Zhou, T. *et al.* Peripheral blood gene expression as a novel genomic biomarker in complicated sarcoidosis. *PloS one* **7**, e44818, doi:10.1371/journal.pone.0044818 (2012).
11. Bloom, C. I. *et al.* Transcriptional blood signatures distinguish pulmonary tuberculosis, pulmonary sarcoidosis, pneumonias and lung cancers. *PloS one* **8**, e70630, doi:10.1371/journal.pone.0070630 PONE-D-13-12284 (2013).
12. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research* **30**, 207–210 (2002).
13. Betel, D., Wilson, M., Gabow, A., Marks, D. S. & Sander, C. The microRNA.org resource: targets and expression. *Nucleic acids research* **36**, D149–153, doi:10.1093/nar/gkm995 (2008).
14. Betel, D., Koppal, A., Agius, P., Sander, C. & Leslie, C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol* **11**, R90, doi:10.1186/gb-2010-11-8-r90gb-2010-11-8-r90 (2010).
15. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25–29, doi:10.1038/75556 (2000).
16. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic acids research* **32**, D277–280, doi:10.1093/nar/gkh063 (2004).
17. Maertzdorf, J. *et al.* Common patterns and disease-related signatures in tuberculosis and sarcoidosis. *Proc Natl Acad Sci USA* **109**, 7853–7858, doi:10.1073/pnas.11210721091121072109 (2012).
18. Sharma, S. M. *et al.* Insights in to the pathogenesis of axial spondyloarthropathy based on gene expression profiles. *Arthritis Res Ther* **11**, R168, doi:10.1186/ar2855ar2855 (2009).
19. Olnes, M. J. *et al.* Effects of Systemically Administered Hydrocortisone on the Human Immunome. *Scientific reports* **6**, 23002, doi:10.1038/srep23002 (2016).
20. Venet, D., Dumont, J. E. & Detours, V. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput Biol* **7**, e1002240, doi:10.1371/journal.pcbi.1002240PCOMPBIOL-D-11-00571 (2011).
21. Wei, C. *et al.* Circulating miRNAs as potential marker for pulmonary hypertension. *PloS one* **8**, e64396, doi:10.1371/journal.pone.0064396PONE-D-12-38535 (2013).
22. Sarrion, I. *et al.* Role of circulating miRNAs as biomarkers in idiopathic pulmonary arterial hypertension: possible relevance of miR-23a. *Oxid Med Cell Longev* **2015**, 792846, doi:10.1155/2015/792846 (2015).
23. Xing, Y. *et al.* MicroRNA-30c contributes to the development of hypoxia pulmonary hypertension by inhibiting platelet-derived growth factor receptor beta expression. *Int J Biochem Cell Biol* **64**, 155–166, doi:10.1016/j.biocel.2015.04.001S1357-2725(15)00102-8 (2015).
24. Hale, A. E., White, K. & Chan, S. Y. Hypoxamirs in Pulmonary Hypertension: Breathing New Life into Pulmonary Vascular Research. *Cardiovasc Diagn Ther* **2**, 200–212, doi:10.3978/j.issn.2223-3652.2012.08.01 (2012).
25. Meloche, J. *et al.* MiR-223 Reverses Experimental Pulmonary Arterial Hypertension. *Am J Physiol Cell Physiol*, ajpcell 00149 02015, doi:10.1152/ajpcell.00149.2015ajpcell.00149.2015 (2015).
26. Cao, M. *et al.* MiR-23a regulates TGF-beta-induced epithelial-mesenchymal transition by targeting E-cadherin in lung cancer cells. *Int J Oncol* **41**, 869–875, doi:10.3892/ijo.2012.1535 (2012).
27. Zhong, K., Chen, K., Han, L. & Li, B. MicroRNA-30b/c inhibits non-small cell lung cancer cell proliferation by targeting Rab18. *BMC Cancer* **14**, 703, doi:10.1186/1471-2407-14-7031471-2407-14-703 (2014).
28. Yang, Y. *et al.* The role of microRNA in human lung squamous cell carcinoma. *Cancer Genet Cytogenet* **200**, 127–133, doi:10.1016/j.cancergencyto.2010.03.014S0165-4608(10)00147-0 (2010).
29. Liang, H. *et al.* MicroRNA-223 delivered by platelet-derived microvesicles promotes lung cancer cell invasion via targeting tumor suppressor EPB41L3. *Mol Cancer* **14**, 58, doi:10.1186/s12943-015-0327-z (2015).
30. Crouser, E. D. *et al.* Differential expression of microRNA and predicted targets in pulmonary sarcoidosis. *Biochem Biophys Res Commun* **417**, 886–891, doi:10.1016/j.bbrc.2011.12.068S0006-291X(11)02277-7 (2012).
31. Muro, S. *et al.* Expression of IL-15 in inflammatory pulmonary diseases. *J Allergy Clin Immunol* **108**, 970-975, doi :S0091-6749(01)74692-110.1067/mai.2001.119556 (2001).
32. Kim, H. S. *et al.* Association of interleukin 23 receptor gene with sarcoidosis. *Dis Markers* **31**, 17–24, doi:10.3233/DMA-2011-0796R05G542244H73436 (2011).
33. Iannuzzi, M. C. *et al.* Genome-wide search for sarcoidosis susceptibility genes in African Americans. *Genes Immun* **6**, 509–518, doi:10.1038/sj.gene.6364235 (2005).
34. Kiszalkiewicz, J. *et al.* Altered miRNA expression in pulmonary sarcoidosis. *BMC medical genetics* **17**, 2, doi:10.1186/s12881-016-0266-6 (2016).

35. Jazwa, A. *et al*. Differential inflammatory microRNA and cytokine expression in pulmonary sarcoidosis. *Arch Immunol Ther Exp (Warsz)* **63**, 139–146, doi:10.1007/s00005-014-0315-9 (2015).
36. Nunes, H. *et al*. Cardiac sarcoidosis. *Seminars in respiratory and critical care medicine* **31**, 428–441 (2010).
37. Zajicek, J. P. *et al*. Central nervous system sarcoidosis–diagnosis and management. *Qjm* **92**, 103–117 (1999).
38. Statement on sarcoidosis. Joint Statement of the American Thoracic Society (ATS), the European Respiratory Society (ERS) and the World Association of Sarcoidosis and Other Granulomatous Disorders (WASOG) adopted by the ATS Board of Directors and by the ERS Executive Committee, February 1999. *Am J Respir Crit Care Med* **160**, 736–755 (1999).
39. Ritchie, M. E. *et al*. A comparison of background correction methods for two-colour microarrays. *Bioinformatics* **23**, 2700–2707, doi:10.1093/bioinformatics/btm412 (2007).
40. Irizarry, R. A. *et al*. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2003).
41. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
42. Affymetrix. *Exon Array Background Correction*, http://media.affymetrix.com/support/technical/whitepapers/exon_background_correction_whitepaper.pdf (2005).
43. Pitroda, S. P. *et al*. Tumor endothelial inflammation predicts clinical outcome in diverse human cancers. *PloS one* **7**, e46104, doi:10.1371/journal.pone.0046104 (2012).
44. Zhou, T., Wang, T. & Garcia, J. G. Genes influenced by the non-muscle isoform of Myosin light chain kinase impact human cancer prognosis. *PloS one* **9**, e94325, doi:10.1371/journal.pone.0094325 (2014).
45. Zhou, T., Wang, T. & Garcia, J. G. Expression of nicotinamide phosphoribosyltransferase-influenced genes predicts recurrence-free survival in lung and breast cancers. *Scientific reports* **4**, 6107, doi:10.1038/srep06107 (2014).

Acknowledgements

This study is supported by National Institutes of Health grant R01HL094394 (JGNG).

Author Contributions

T.Z., N.C. and J.G.N.G. conceived of the study. T.Z., N.C., K.S.K. and J.G.N.G. participated in the design of the study. T.Z. and N.C. collected the microarray data. T.Z. processed the microarray data. T.Z. performed the statistical analysis. T.W., Y.L., K.S.K. and J.G.N.G. helped interpret the results. T.Z., N.C., T.W., Y.L., K.S.K. and J.G.N.G. drafted the manuscript. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-04109-6

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.